Abstract

Despite growing evidence for learning loss due to COVID-19, there is little research examining this phenomenon using norm-referenced tests (NRTs) or with special educations students. Using a repeated-measures design with 96 fourth through 12th grade students previously identified as eligible for special education services, the present study attempted to measure learning loss using *W* Difference Scores from the Woodcock Johnson IV Tests of Achievement. Findings revealed statistically significant learning loss in reading decoding, spelling, and math calculation skills. Academic proficiency was found to differ markedly from normative expectations for typical same-age peers across tests, both prior to and during the COVID-19 pandemic. While academic proficiency was more similar to that of a clinical sample of students with learning disabilities, moderate to large differences in the sample means of most tests suggest that COVID-19 has had a negative impact on academic achievement. Implications regarding practice and future research are discussed.

Keywords: Norm-referenced tests; COVID-19; learning loss; repeated measures; academic achievement

A Comparison of Special Education Students' Norm-Referenced Academic Achievement Before and During COVID-19

In spring 2020, COVID-19 spread rapidly across the United States precipitating school closures (Donohue & Miller, 2020; Viner et al., 2020) as well as economic hardship (Center for Budget and Policy, 2021). While most school personnel were not prepared for the rapid transition from in-person to remote instruction, the flexibility of remote approaches was generally well-received. However, concerns regarding restricted interaction, inadequate infrastructure (e.g., power, internet access), and equipment (e.g., lack of access to laptops) were significant during the transition (Hebebci et al., 2020).

In a technical report by McKinsey and Company, the authors expressed concerns regarding potential learning loss as a result of COVID-19 restrictions and noted that (a) remote students were observed engaging in less learning activities compared to when schools were open and that (b) students may have been less externally motivated to engage in learning tasks in a remote learning environment (Dorn et al., 2020). In addition, the authors predicted that stress related to COVID-19 would affect concentration and learning efforts. Finally, students from less-advantaged homes were believed to be more likely impacted by infrastructure gaps and learning inequity (Dorn et al., 2020).

Benchmark Tests

Learning loss related to COVID-19 has since been reported in several large studies conducted in the United States examining benchmark test scores. NWEA (i.e., Kuhfeld et al., 2020) released a research brief examining data from 4.4 million U.S. students who completed MAP Growth™ assessments during the fall of 2020. These researchers found that while 3-8 grade students performed similarly to their 2019 same-grade peer in reading, their scores in math

fell between 5 to 10 percentile points below the fall 2019 normative sample (Kuhfeld et al., 2020).

Renaissance Learning (2021) published an executive summary examining the adaptive Star Assessments scores of roughly 3.8 million children in grades 1-8 collected during the middle of the 2020-2021 school year. These researchers found that on average, students scored slightly (two percentile points) below pre-covid mid-term expectations in reading and six percentile points below in math. In addition to significant learning loss in math, English learners, students with disabilities, and students who attended urban and Title 1 schools were significantly more likely to have scores that were negatively impacted. Interestingly, students from rural schools saw an improvement in both math and reading (Renaissance Learning, 2021).

A research Brief by Curriculum Associates (2021) examined i-Ready Diagnostic data collected from over nine million students. Findings from this study suggested that fewer students were able to access grade-level work a year after COVID-19 first interrupted instruction. These losses were less pronounced in reading, especially in grades 1-6, which saw a 2-10% drop in the percentage of students who were on grade level during Winter 2021 assessment. The percentage of students who were on grade level during Winter 2021 assessment in mathematics decreased from between 8% and 16% in students, grade 1-6. Furthermore, these researchers indicated that learning loss was greatest in schools that serve majority Black and Latinx students and in schools located in lower-income zip codes (Curriculum Associates, 2021).

Additionally, an executive summary by Istation (Patarapichayatham et al., 2021) examined Istation's Indicators of Progress (ISIP™) assessments reading data from over 290,000 students and math scores from 34,000 pupils. Their results suggested that students experienced roughly two months of learning loss in reading between April 2020 and May 2021. Losses were

greater for math equating to approximately one to three months of learning loss for younger students and four to five months for students in upper grades. All these losses were in addition to the expected one to two months of learning loss due to summer vacation (Patarapichayatham et al., 2021). In sum, all of these studies suggested that students have experienced significant loss in math and most found slight losses in reading. Additionally, results seemed to indicate that students with disabilities, from economically disadvantaged families and from minoritized groups have been disproportionately impacted by COVID-19.

Norm-referenced Tests

While the above studies focused on benchmark testing, scant literature has examined whether learning loss has been detected using norm-referenced tests (NRTs). NRTs are high stakes because scores obtained from them are heavily relied upon to determine eligibility for special education services (Lockwood et al., 2021). Furthermore, because these tests are used to examine the progress of students who are suspected of having a disability, they provide data for students who are struggling academically. A comprehensive search of the literature revealed only two studies examining the impact of COVID-19 on NRTs of academic achievement. A study conducted by researchers at Pearson (Raiford et al., 2021) found no evidence of learning loss in their analysis of data collected via Pearson's Q-interactive digital platform. Specifically, the authors compared group averages from students tested with the Kaufman Test of Educational Achievement, 3rd Edition (KTEA-3; Kaufman & Kaufman, 2014) and the Wechsler Individual Achievement Test 3rd Edition (WIAT-III; Wechsler, 2009) in spring and summer of 2019 to group averages in spring and summer of 2020. The sample sizes for the KTEA-3 were 9,334 in 2019 and 3,030 in 2020. The sample sizes for the WIAT-III were 9,587 in 2019 and 4,480 in 2020. While small differences were noted in some composite scores, the authors noted that

academic achievement scores were “highly consistent” (Raiford et al., 2021, p. 2) across the 2019 and 2020 samples.

There are considerable limitations with this study. Raiford et al., (2021) noted that this research was largely ad hoc and based on data collected by practitioners using Pearson’s digital platform. Due to this, participant matching was not completed. Additionally, while it is true that scores were largely unaffected, there are several explanations. The first is that students did not experience learning loss during COVID-19. While it is possible that the students in the sample examined by Raiford and colleagues (2021) did not experience learning loss, this seems unlikely as multiple studies (i.e., Curriculum Associates, 2021; Kuhfeld et al., 2020; Patarapichayatham et al., 2021; Renaissance Learning, 2021) using benchmark test data suggest otherwise, particularly in math. Another possibility is that losses did occur but that they were not detected by Raiford et al. As noted by the authors, their study used convenience sampling and should therefore be “considered preliminary” (p. 18). Additionally, pandemic scores were obtained from May to August 2020 (Raiford et al., 2021), which was early in the pandemic.

Another study by Lupas and colleagues (2021) examined the extent to which learning loss was experienced by 116 students as measured by three subtests (i.e., Word Reading, Spelling, and Numerical Operations) of the WIAT-III. These researchers found that scores significantly increased on a test of reading and spelling; no significant changes were detected in math. However, there were multiple limitations of this study. First, the sample was restricted to students with attention deficit hyperactivity disorder (ADHD) some of whom started medication treatment in between pre- and posttests. Second, posttest scores were obtained after only three to four months of remote instruction, which was early into the pandemic. Third, fall test scores were obtained during the summer of 2019 and Lupas et al. (2021) noted that “the fall assessment

may have deflated the beginning-of-year scores artificially" and therefore increases "seen in scores may be an artifact" of this confound (p. 321). Finally, posttest scores were obtained remotely, violating standardization (Lupas et al., 2021). This violation of standardization in administration may have introduced a number of construct-irrelevant variables (Farmer et al., 2020) and, potentially, unreliable data (Gilbert et al., 2021).

Purpose

The purpose of this study was to examine differences in scores on the Woodcock-Johnson IV Tests of Achievement (WJ IV ACH; Schrank et al., 2014), one of the most commonly administered omnibus NRT of academic achievement by school personnel (Benson et al., 2019; Lockwood et al., 2021), before and during COVID-19. This examination is important as researchers (e.g., Kuhfeld et al., 2020; Renaissance Learning, 2021) have suggested that underachieving students, especially those served in special education, may have experienced more learning loss compared to their general education peers. Examining omnibus NRT scores is ideal because these tests are used almost exclusively with underachieving students and students with disabilities in educational settings and are used to make high-stakes decisions. As noted previously, Raiford and colleagues (2021) and Lupas et al.'s (2021) failure to detect any learning loss may be an artifact of their sampling or other methodological issues. In contrast to the research by Raiford et al. (2021), we used a repeated-measures design to provide an experimental control for individual differences that may have obscured the effect of the COVID-19 on test scores. Additionally, as we collected scores from administrations ranging from August of 2020 until April of 2021, our results may better measure the full impact of COVID-19 on achievement than any of the mentioned studies. Furthermore, unlike Lupas et al., (2021) we wanted to examine students across categories (not just ADHD), and to address other previously noted

confounds in their study. We posited the following hypothesis, which was pre-registered on 10/28/2020 (https://osf.io/2bmjg/?view_only=9128ca5dec3f4fd8ae265c7f3cca45e3):

There will be a significant decrease in observed scores on the core subtests of the WJ IV ACH when comparing tests conducted pre-COVID-19 to administrations after April 2020.

Methods

Participants

This study was approved by the lead author's, *and* the district's institutional review boards. To protect the participants' anonymity, all materials were collected *and* de-identified by a district employee before being sent to the researchers. De-identified test performance data were obtained for 96 participants who (a) had been tested prior to the pandemic using the WJ IV ACH and subsequently found eligible for special education services and (b) were reevaluated during the pandemic using the WJ IV ACH. At the time of the reevaluations, participants were sampled from 38 schools; 27 of these were elementary and 11 were secondary. Most referrals for testing were related to specific learning disabilities (70.9%), followed by other health impairments (10.4%), emotional disturbances (5.2%), autism spectrum disorder (6.3%), specific language impairments (4.2%), intellectual disability (2.1%), and traumatic brain injury (1.0%). Most of the students tested were male (60.4%) and Latinx (57.3%), with 32.3%, 5.2%, 2.1%, and 2.1% of those tested reported as being White, Black, Native American, and Multiracial, respectively. The grade level of participants at first testing ranged from kindergarten to ninth (M = 3.32, SD = 2.02) and from fourth to twelfth at second testing (M = 6.22, SD = 2.03). More information regarding grade and dates of testing can be found in Table 1 and in an OSF Supplemental Table (https://osf.io/v9537/?view_only=a3b76a43b1a948f6b011acca9446b48f).

Instrument

Test record forms from administrations of WJ IV ACH were reviewed in this study. Scores from the following WJ IV ACH clusters were used when comparing pre-pandemic achievement to achievement during the pandemic: Basic Reading Skills, Reading Fluency, Reading Comprehension, Math Calculation Skills, Math Problem Solving, and Written Expression. Extensive validity evidence is available to support interpretations of scores derived from the WJ IV ACH (McGrew et al. 2014; Niileksela et al., 2016). Difference curves across age levels indicate rapid acceleration of growth in the academic skills included in this study from age six to about age 15. Cluster reliabilities exceed .90 at all relevant age levels, and cluster intercorrelations are higher for clusters from the same achievement domain relative to clusters from different achievement domains (McGrew et al., 2014).

The metrics of analysis in this study were *W* scores and *W* Difference scores. *W* scores (Woodcock & Dahl, 1971) are a transformation of ability and item scores from a Rasch measurement model and are unique in that they are on an equal-interval scale (i.e., a given interval, such as 10 points, has the same meaning across the scale and across the academic skills measured). The equal interval feature is critical in studies that examine changes in scores over time (Embretson, 1996; McArdle et al., 2002; McDonald, 1999). *W* scores have been used in longitudinal analyses that examined change in cognitive abilities (McArdle et al., 2002; Tucker-Drob, 2009). Academic skills tend to increase as a result of educational experiences, thus *W* scores for school-aged children and adolescents tend to increase over time. *W* Difference scores represent the distance of an examinee's *W* score from a grade-appropriate reference point (i.e., average scores at the examinee's grade level based on WJ IV normative data). We analyzed *W* Difference scores in the present study because this metric provides a pre-pandemic, normative proficiency criteria against which the samples' current academic achievement can be compared.

Although standard scores are a more familiar metric for most readers, these linear transformations of raw scores have properties that limit their usefulness as measures of growth. Test scores are "…intended to locate a person on the variable defined by the test items taken" (Wright & Stone, 1979, p. 4). It is commonly assumed that standard scores adequately locate individuals on measured variables and thus are useful for measuring growth. Standard scores indicate differences, in standard deviation units, between raw scores and the mean of the comparison group. Although standard scores accurately estimate rank order within a comparison group, spacing between rankings is imprecise. Consequently, when standard scores are used to examine growth, the results will be distorted and possibly misleading (Wright & Stone, 1979).

In Rasch models, the probability of a correct response is dependent on the difference between the ability of the person and the difficulty of the item (Rasch, 1960). The probability of answering any item correctly increases as ability increases, while the probability that any person can answer an item correctly increases as the easiness of an item increases. The location of individuals on the Rasch scale are determined by their performance on the test items taken, and these locations represent individuals' ability level (*B*) in the variable defined by the test items. The basis for the W scores used in our analyses is *B*, which is transformed by changing the units of *B* using a logarithm with base 9, then multiplying by 20 and rounding to improve interpretability, and finally adding a constant of 500 to reduce the likelihood of negative values (Benson et al., 2018).

The *W* scale represents proficiency (i.e., how much a student can do in a given academic domain) and can be interpreted across age- and grade-level. *W* difference scores are the simple difference between an examinee's obtained *W* score and a reference point (i.e., the median score for normative data at the examinee's grade level). Larger *W* difference scores indicate greater

distance between an examinee's *W* score and the reference point. Positive *W* difference scores indicate that the examinee's *W* score is above average relative to same-grade peers while negative *W* difference scores indicate that the examinee's *W* score is below the normative average. In contrast to standard scores, which indicate relative rank, *W* difference scores reflect degree of proficiency on tasks mastered by peers at the examinee's grade level.

Procedure

Approved employees from a large Southwestern school district scanned and submitted two de-identified WJ IV ACH test record forms from administrations to the same child as part of the special education evaluation process; one administration occurred pre-pandemic and the other administration occurred during the 2020-2021 school year. All students that were tested using the WJ IV ACH during the 2021-2022 school year who also had a pre-COVID-19 WJ IV ACH score report located in the districts' database were sampled for this study. We received 96 pairs of pre- and post-pandemic record forms, with an average of 36 months (SD = 3.37) between administrations. Pre-pandemic testing dates range from January of 2017 to April of 2019. Testing dates during the pandemic range from August of 2020 to April of 2021. Data were entered into SPSS by two graduate student coders and double checked for redundancy and accuracy of data. Percent of agreement of coding was 99%. All discrepancies were discussed by the research assistants and a consensus reached regarding correct coding. Our dataset is open source and can be accessed at https://osf.io/v9537/?view_only=a3b76a43b1a948f6b011acca9446b48f.

Setting

The WJ IV ACH was administered according to district protocol which, during COVID-19, included the use of a clear plastic partition and six feet of social distancing when possible.

Testing to determine special education eligibility was halted in March 2020 and fully resumed in August 2020 for K-12 students. All testing occurred in person on school grounds. All district schools moved to remote instruction mid-March 2020 through the end of the 2020 school year. Remote instruction continued the beginning of the 2020-2021 school year, with the exception of self-contained classrooms that were made available when school started. Support centers were established in the middle of August 2020 where students could engage in online instruction under the supervision of adults, with a staff to student ratio that ranged from 2:9 to 1:20. The support centers were available for select students including students with disabilities and those that lacked necessary technology to participate in remote learning. In-person instruction became an option for students the middle of October 2020 and remained so for the entire 2020-2021 school year with the exception of a two-week move back to remote instruction during January. Remote, synchronous instruction was also provided as an option for all students. This included special educations students who had the option of receiving resource, small group, and self-contained services "remotely." Roughly 33% of special education students enrolled in remote instruction throughout the 2020-2021 school year, compared to approximately 36% of general education students. Academic interventions continued throughout the pandemic; of note, interventions were provided at the discretion of individual teachers, schools, and/or special education teams and no consistent interventions or progress monitoring were used across the district.

According to the most recent (2017-2018) fiscal data, the districts' locale was characterized as City: Large with over 60,000 students. Roughly 17% of students in the district had an individualized education program. The student population was majority white (61%),

followed by Latinx (28%). Approximately 20% of students' families had an income under the poverty line (National Center for Education Statistics, n.d.).

Data Analysis

A comparison sample was created using mean *W* scores and standard deviations from Appendix Table B-3 of the WJ IV Technical Manual (McGrew et al., 2014) to control for possible "Matthew Effects" (i.e., the widening of achievement gaps between disabled and non-disabled children over time; Scarborough & Parker, 2003). These means and standard deviations were multiplied by the number of participants at each age level, then products for each test were summed and divided by the sample size for that test. Separate comparison samples were created to correspond to the ages of participants when tested before and during the pandemic. As our sample consists of students with disabilities, we also created comparison groups using data from the WJ IV Clinical Validity Samples. The publisher of the WJ IV was contacted and provided these data for use in this study. Participants were assigned to younger and older age groups to match the average age at testing for our data collected before and during the pandemic. This resulted in a total sample of 142, with 71 assigned to the younger group and 71 assigned to the older group. Sample sizes, as well as *W* score means and standard deviations for the study sample and comparison samples, are presented in Table 2.

Univariate and multivariate measures of skew and kurtosis were examined to check for violations of normality in the score distributions of *W* Difference scores. Estimates of univariate kurtosis for Writing Samples, Word Attack, and Letter-Word Identification are 14.26, 5.49, and 3.02, respectively, in data obtained during the pandemic. Elevated kurtosis may result from issues such as regression to the mean or a trend toward more targeted assessment during reevaluations so that practitioners are less likely to administer these tests to students with scores

toward the extremes of the distribution relative to students with scores near the center of the distribution. Likewise, an estimate of -3.29 indicates problematic skew in the distribution of Writing Samples data during the pandemic, which indicates possible regression to the mean for extremely low scores and/or practitioners being less likely to administer this test to students who are likely to score in the lower tail of the distribution. Omnibus tests of multivariate normality (see Looney, 1995) were conducted across pairs of variables representing performance before and during the pandemic. The hypothesis that data are multivariate normal was not supported for Calculation, Letter-Word Identification, Passage Comprehension, Spelling, Word Attack, or Writing Samples (p values < .001). Multivariate normality was supported for Applied Problems (p = .81) and Spelling (p = .08).

Missing data were present because administrators were allowed by the district to use their educational judgment to determine which subtests were appropriate to administer based on each student's referral concern. Missing data analysis revealed that about 3% of values are missing from pre-pandemic testing and about 6% are missing from testing that occurred during the pandemic. Spelling had the most missing values, with about 10% of values missing from pre-pandemic testing and about 24% missing from testing that occurred during the pandemic. Excluding Spelling, only 2% of values are missing from pre-pandemic testing and only 3% of values are missing from testing that occurred during the pandemic. Analysis of missing value patterns did not reveal any patterns of missingness that occurred in more than 5% of the sample. However, it is unlikely that data are missing completely at random, but they are missing for a predictable reason (i.e., are missing at random based on referral questions at initial assessment or assessment needs at reevaluation) rather than missing due to testees' levels on these variables (i.e., missing not at random). Multiple imputation was selected as an appropriate method for

handling missing data (Enders, 2017). Automatic imputation was conducted using SPSS Version 28 in a linear regression model with 50 iterations.

Paired-samples t-tests were used when analyzing Applied Problems and Spelling. A two-tailed test of significance was applied to the pooled data (i.e., all imputed data sets were combined), with the significance level set to 1%. As multivariate normality was not supported for other variables, all other score comparisons were made using a nonparametric test, the Wilcoxon signed ranks test. The Wilcoxon signed ranks test compares ranks of scores from matched pairs, in this case pairs of scores from the same individual sampled before and during the pandemic, to test the null hypothesis that the difference between pairs follows a symmetric distribution around zero. Two-tailed tests of significance were tested with the two-tailed significance level set to 1%. Exact significance tests were used to maintain the Type I error rate at the desired significance level. Results from the original data set are reported, although all imputed data sets were examined to ensure that *p* values obtained using these data were consistent with *p* values obtained using the original data. Pooling methodology is based on Rubin's (2004) rules, which do not address nonparametric tests, so we set the conservative a priori criterion that statistically significant findings should be evidenced by *p* values that consistently fall below .01 across all data sets. The sample size was sufficiently large to detect effect sizes of .40 or higher with at least 95% power and effect sizes of .30 or higher with at least 80% power.

Standardized effect sizes are reported for the comparisons conducted in this study. Corrections to Cohen's *d* (Cohen, 1988) were used to calculate effects representing between-sample differences between the study sample and comparison samples derived from WJ IV normative and WJ IV clinical validity data. Specifically, Hedge's *g* and Glass's delta, were used

to account for unequal samples sizes and unequal sample means, respectively. SPSS provides *Z* values from the Wilcoxon signed ranks test used for paired comparisons. *Z* values were converted to effect size estimates using the following formula provided by Rosenthal (1994):

$$r = \frac{Z}{\sqrt{N}}$$

Cohen's *d* was used to calculate the effect size for the Math Calculation paired-samples t-test. Additionally, we report average W Difference scores and the Relative Proficiency Index (RPI) to facilitate interpretation of performance on the *W* scale. The RPI reflects the relative proficiency of the examinee at the difficulty level at which peers were 90% proficient. The average range for this index is 82 to 95. Scores ranging from 67 to 82 are indicative of limited to average proficiency. Scores ranging from 24 to 67 indicate limited proficiency and scores below 24 indicate that proficiency is very limited to extremely limited.

Results

W scores for the study sample and comparison samples are presented in Table 2. As expected, *W* scores were found to increase as students receive academic instruction and mature developmentally. *W* scores for all tests increased from the first to the second measurement occasion for the study, clinical, and normative samples. *W* scores also increased for both the clinical sample and normative sample when comparing samples with a mean age of 9 to samples with a mean age of 12. Notably, *W* scores for the study sample and the clinical sample are considerably lower than *W* scores for same-age peers in the normative sample. With the exception of Word Attack, *W* scores for the study sample are lower than those for the clinical sample.

Standardized effect sizes for *W* score comparisons between the WJ IV Clinical, WJ IV Normative, and study samples are presented in Table 3. The largest average effects were observed for the comparison between the study sample and the normative sample, with means for the normative sample being more than a standard deviation higher for four of seven tests at the first measurement occasion and six of seven tests at the second measurement occasion. For the remaining comparisons, means for the normative sample are more than a half standard deviation higher. Results indicate a time effect wherein the average difference in test means increased from 1.16 to 1.34, indicating that the gap widened over time so that the study sample fell further behind expectations based on nationally representative test norms. Standardized effect sizes for *W* score comparisons between the study sample and clinical sample are smaller than those for comparisons with the normative sample across comparisons. Notably, the mean for Word Attack and Writing Samples obtained on the second measurement occasion in the study sample exceed comparable means from the clinical sample. However, means for the other 12 comparisons are consistently lower for the study sample, with most of these differences being moderate to large in magnitude.

Mean *W* Difference scores and mean RPI scores are presented in Table 4. Average *W* Difference scores and average RPIs indicate that students in the sample fell further behind expectations with respect to skills measured by Applied Problems, Calculation, Spelling, and Word Attack. The decline in proficiency with Applied Problems, $t = 1.72$, $p = .09$, $d = .18$, 95% CI [-.03, .39], was not statistically significant and small in magnitude. The decline in proficiency with Calculation, $Z = -3.61$, $p = < .001$, $r = -.41$, was statistically significant and moderate in magnitude. The decline in proficiency with Spelling, $t = 3.79$, $p = < .001$, $d = .47$, 95% CI [.21, .72], was also statistically significant and moderate in magnitude. The decline with Word Attack,

$Z = -1.98$, $p = .05$ $r = -.87$, was statistically significant and large in magnitude. Wilcoxon signed ranks tests revealed statistically significant improvement for Letter-Word Identification, $Z = -2.87$, $p = < .01$, $r = -.30$, and Writing Samples, $Z = -2.61$, $p = <.01$, $r = -.25$, but not for Passage Comprehension, $Z = -2.26$, $p = .012$, $r = -.28$, although all of these effects are relatively small and similar in magnitude.

Discussion

Multiple studies (i.e., Curriculum Associates, 2021; Kuhfeld et al., 2020; Patarapichayatham et al., 2021; Renaissance Learning, 2021) have examined learning loss using benchmark data. All of these studies suggested that students experienced learning loss in math. Of note, none of these studies examined writing and none were peer-reviewed. Only two publications examined learning loss using NRTs (i.e., Lupas et al., 2021; Raiford et al., 2021) and suggested that no learning loss occurred as measured by two of the most commonly administered omnibus measures of academic achievement (i.e., KTEA-3, WIAT-3; Benson et al., 2019; Lockwood, 2021). However, Raiford et al.'s (2021) paper was also not peer reviewed and the Lupas et al. (2021) study solely examined students with ADHD. Ours is the first peer-reviewed study to examine learning loss experienced by a solely special education student population across disability categories, and the only to study the extent to which the WJ IV ACH detected changes in learning due to COVID-19.

Our sample's math proficiency was limited before and during COVID-19. Like the benchmark studies (i.e., Curriculum Associates, 2021; Kuhfeld et al., 2020; Patarapichayatham et al., 2021; Renaissance Learning, 2021) our sample experienced significant learning losses in math calculation (i.e., Calculations). Proficiency with math problem solving (i.e., Applied Problems) also decreased, although the observed learning loss was not statistically significant.

As the preponderance of studies using benchmark data (i.e., Curriculum Associates, 2021; Kuhfeld et al., 2020; Patarapichayatham et al., 2021; Renaissance Learning, 2021) found similar results, we believe that it is likely that students have experienced significant learning loss in math due to COVID-19. However, Raiford and colleagues (2021) and Lupas et al. (2021) did not find differences in their study using NRTs. We believe that there are several explanations for this. One is due to Raiford et al.'s lack of a repeated-measures design, which we used as it (a) better controls for confounds and (b) provides increased power to detect changes. Additionally, while Lupas et al. (2021) used a repeated-measures design, multiple confounds were noted including deflated pre-test scores, and the violation of standardization rules in posttest administration that may have obfuscated learning losses.

Another plausible explanation is our use of *W* Difference scores. Standard scores indicate relative rank of performance within the range of scores obtained by peers while *W* Difference scores indicate degree of proficiency on criterion tasks mastered by average peers. It is more meaningful to compare changes in distance from a grade-appropriate reference point on the *W* scale than it is to compare changes in relative rank, as changes in relative rank can occur due to individual differences in performance that are unrelated to proficiency status. Thus, *W* Difference scores provide a more accurate measure of proficiency status and are more sensitive to intraindividual change relative to standard scores. Finally, our test scores were obtained further into the pandemic (both Lupas et al. and Raiford et al. collected data in spring or summer 2020) and may better capture the true extent of the learning loss that students experienced due to COVID-19. Regardless of the explanation, the preponderance of evidence from our study as well as Renaissance Learning (2021) and Kuhfeld et al. (2020) suggests that students with disabilities

have experienced significant learning loss in math beyond what would be expected had these students' instruction not been interrupted by COVID-19.

Results indicate that our sample experienced statistically significant learning loss in word decoding skills (i.e., Word Attack), but minimal learning loss to modest improvement in sight-word identification (i.e., Letter-Word Identification) and comprehension (i.e., Passage Comprehension). This is fairly consistent with other studies which suggest negligible loss (i.e., Kuhfeld et al., 2020; Raiford, 2021) or minimal learning loss (i.e., Curriculum Associates, 2021; Renaissance Learning, 2021) in reading. We can only attribute this to the continuation of reading interventions throughout the pandemic. It is possible that most students in our sample were receiving effective reading instruction and were able to maintain learning gains despite instructional disruptions during the pandemic. However, it is of note that proficiency with reading skills remained in the limited range for our sample throughout the pandemic. This suggests that while some gains may have been made, students in our sample continued to underperform compared to their same-grade peers.

Results for written expression were mixed, suggesting a moderate decline in proficiency with spelling (i.e., Spelling) and a small increase in writing proficiency (i.e., Writing Samples). These relatively small effects are consistent with Raiford et al. (2021), who found no significant differences in written expression on either the KTEA-3 or WIAT-3. It is worth noting that the students in our sample were relatively proficient in written expression before COVID-19 compared to all other subject areas.

Implications for Research and Practice

Like all the previously mentioned benchmark testing studies (i.e., Curriculum Associates, 2021; Kuhfeld et al., 2020; Patarapichayatham et al., 2021; Renaissance Learning, 2021), our

sample experienced significant learning loss in math calculation. This indicates that students with disabilities may need intensive intervention in math above and beyond the interventions that they would already receive in tiers 1-3. Because appropriate math interventions vary depending on the grade and specific deficit area of students, we highly encourage educators to consult with their districts' math interventionists or other authoritative sources to find evidence-based interventions for their students. We also found declines in spelling and decoding; educators are highly encouraged to conduct progress monitoring to determine their students' needs and to target interventions accordingly in spelling and decoding.

As previously noted, the average time between test administrations was three years. Due to this long interval, which stretched longer than we have experienced instructional disruptions, it is possible that learning losses that students experienced due to COVID-19 were not fully detected as they were offset by gains that were made due to interventions provided prior to COVID-19—or that losses detected were not solely due to COVID-19 related instructional disruption. Additionally, it is possible that interruptions in instruction will continue next year due to new variants of the virus. Future research examining differences in scores obtained just prior to COVID-19 and student performance until the pandemic has fully abated is recommended; we are unable to rule out historical effects due to the gap observed between the first and second administrations of the WJ IV ACH and concede that the differences identified in our analyses may reflect causal mechanisms beyond COVID-19 and associated changes to instruction. However, the WJ IV normative data provide a reasonable estimate of expected growth during a three-year interval, and data from WJ IV Clinical Validity Studies provide a reasonable estimate of expected growth for students with disabilities. Both the clinical sample and the study sample are comprised of students identified with disabilities, and differences between these two groups

suggest that instructional changes resulting from COVID-19 had an appreciable impact on some academic skill areas.

As our findings are from one sample in a single school district, our findings may, at least in part, be an artifact of our sample and may not generalize to students in other districts. This is especially true because instruction during COVID-19 has varied significantly from state to state and between districts. Because approximately 67% of our sample attended class in person, they may have experienced significantly less learning loss compared to districts that were fully remote the entire school year. Furthermore, no consistent response to intervention (RtI), progress monitoring or interventions were used in this district. Instead, schools or teachers determined what progress monitoring and interventions to use. It may be that students in districts with more coordinated RtI may have experienced less learning loss. Additionally, other studies of benchmark test scores (e.g, Kuhfeld et al., 2020) have found that learning losses varied based on grade. Due to our relatively small sample size, we were unable to examine learning loss by grade level. Relatedly, as we examined reevaluation data, we were unable to examine learning loss experienced by students in grades K-3. Future studies should consider examining large data sets from districts across the United States to examine how learning loss varies by state, district, ethnicity and grade. However, as this is the first peer-reviewed study of learning losses experienced by special education students, across eligibility categories during COVID-19, we believe that this literature provides an important starting place for researchers and educators.

References

Benson, N. F., Beaujean, A. A., Donohue, A., & Ward, E. (2018). W scores: Background and derivation. *Journal of Psychoeducational Assessment*, *36*(3), 273–277. https://doi.org/10.1177/0734282916677433

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology, 72*, 29-48. https://doi.org/10.1016/j.jsp.2018.12.004

Center for Budget and Policy (2021, August). *Tracking the COVID-19 recession's effects on food, housing, and employment hardships*. https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-recessions-effects-on-food-housing-and

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Hillsdale: Lawrence Erlbaum.

Curriculum Associates. (2021). What we've learned about unfinished learning. https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/iready-understanding-student-needs-paper-winter-results-2021.pdf

Donohue, J. M., & Miller, E. (2020). COVID-19 and school closures. *JAMA, 324*(9), 845-847. https://doi.org/10.1001/jama.2020.13092

Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020, December 8). *COVID-19 and learning loss—disparities grow and students need help*. McKinsey & Company. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-learning-loss-disparities-grow-and-students-need-help

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 4,* 341–349.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, *98*, 4–18.

Farmer, R. L., McGill, R. J., Dombrowski, S. C., McClain, M. B., Harris, B., Lockwood, A. B., Powell, S. L., Pynn, C., Smith-Kellen, S., Loethen, E., Benson, N. F., & Stinnett, T. A. (2020). Teleassessment with children and adolescents during the coronavirus (COVID-19) pandemic and beyond: Practice and policy implications. *Professional Psychology: Research and Practice, 51*(5), 477–487. https://doi.org/10.1037/pro0000349

Gilbert, K., Kranzler, J. H., & Benson, N. (2021). An independent examination of the equivalence of the standard and digital administration formats of the Wechsler Intelligence Scale for Children, Fifth Edition. *Journal of School Psychology, 85*. 113-124. https://doi.org/10.1016/j.jsp.2021.01.002

Hebebci, M. T., Bertiz, Y., & Alan, S. (2020). Investigation of views of students and teachers on distance education practices during the Coronavirus (COVID-19) pandemic. *International Journal of Teaching in Education and Sciences, 4*(4), 267-282.

Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement, third edition (KTEA-3)*. Pearson.

Kuhfeld, M., Tarasawa, B., Johnson, A., Ruzek, E., & Lewis, K. (2020). *Learning during COVID-19: Initial findings on students' reading and math achievement and growth.* NWEA Research. https://www.ewa.org/sites/main/files/file-attachments/learning_during_covid-19_brief_nwea_nov2020_final.pdf?1606835922

Lockwood, A. B., Farmer, R., Bohan, K., Winans, S., & Sealander, K. (2021). Academic achievement test use and assessment practices: A national survey of special education

administrators. *Journal of Psychoeducational Assessment*. *39*(4), 436-451. https://doi.org/10.1177/0734282920984290

Lupas, K. K., Mavrakis, A., Altszuler, A., Tower, D., Gnagy, E., MacPhee, F., … & Pelham Jr, W. (2021). The short-term impact of remote instruction on achievement in children with ADHD during the COVID-19 pandemic. *School Psychology*, *36*(5), 313.

Mardia, K. V. (1970): Measures of multivariate skewness and kurtosis with applications. Biometrika, 57, 519-530.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual: Woodcock-Johnson IV*. Riverside.

National Center for Education Statistics. (n.d.) *Education demographic and geographic estimates*. https://nces.ed.gov/Programs/Edge/ACSDashboard/

Niileksela, C. R., Reynolds, M. R., Keith, T. Z., & McGrew, K. S. (2016). A special validity study of the Woodcock-Johnson IV: Acting on evidence for specific abilities. In D. P. Flanagan & V. C. Alfonso (Eds.), *WJ IV clinical use and interpretation: Scientist-practitioner perspectives*. (pp. 65–106). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-802076-0.00003-7

Patarapichayatham, C., Locke, V. N., & Lewis, S. (2021). COVID-19 learning loss in Texas. https://www.istation.com/Content/downloads/studies/COVID-19_Learning_Loss_Texas.pdf

Raiford, S. E., Breaux, K., Chen, S., & Matta, T. (2021). "COVID slide" not evident in individually administered clinical test scores obtained from a large, referred sample. *Pearson.*

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Renaissance Learning. (2021). How kids are performing: Tracking the impact of COVID-19 on reading and mathematics achievement. https://www.renaissance.com/how-kids-are-performing

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). Russell Sage Foundation.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley and Sons.

Scarborough, H. S., & Parker, J. D. (2003). Matthew effects in children with learning disabilities: Development of reading, IQ, and psychosocial problems from grade 2 to grade 8. *Annals of Dyslexia, 53*(1), 47-71.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement*. Riverside.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*(4), 1097–1118. https://doi.org/10.1037/a0015864.supp

Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., Mytton, O., Bonell, C., & Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: A rapid systematic review. *The Lancet Child & Adolescent Health, 4*(5), 397-404. https://doi.org/10.1016/S2352-4642(20)30095-X

Wechsler, D. (2009). *Wechsler individual achievement test* (3rd ed.). Pearson.

Woodcock, R. W., & Dahl, M. N. (1971). *A common scale for the measurement of person ability and item difficulty* (AGS Paper No. 10). American Guidance Service.

Wright, B. D., and Stone, G. (1979). *Best Test Design*. Chicago: MESA Press.

Table 1

Grade of Students at the Time of Testing

	Pre-Pandemic			During Pandemic		
Grade	**N**	**%**	**Cumulative %**	**N**	**%**	**Cumulative %**
K	3	3.1	3.1	0	0.0	--
1	13	13.5	16.7	0	0.0	--
2	37	38.5	55.2	0	0.0	--
3	18	18.8	74	0	0.0	--
4	8	8.3	82.3	15	15.6	15.6
5	4	4.2	86.5	40	41.7	57.3
6	6	6.3	92.7	17	17.7	75.0
7	1	1.0	93.8	6	6.3	81.3
8	2	2.1	95.8	4	4.2	85.4
9	4	4.2	100.0	8	8.3	93.8
10	0	0.0	--	0	0.0	93.8
11	0	0.0	--	2	2.1	95.8
12	0	0.0	--	4	4.2	100.0

Note. Percentages have been rounded to the nearest tenth.

Table 2

Descriptive Statistics for W Scores Obtained from the for Woodcock-Johnson IV (WJ IV) Tests of Achievement

			Study Sample			WJ IV Clinical Sample			WJ IV Normative Sample		
		Age		*W Score*			*W Score*			*W Score*	
Test	Measurement/ Age Grouping	*M*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Word Attack	First	9	94	470.4	22.3	34	470.7	20.4	96	485.3	17.8
Word Attack	Second	12	91	482.8	18.1	31	477.7	11.8	96	500.9	15.6
Letter-Word ID	First	9	96	432.4	42.6	71	452.5	43	96	474.4	27.6
Letter-Word ID	Second	12	94	469.9	33.8	71	479.5	26.9	96	507	23.0
Passage Comprehension	First	9	96	444.6	29.7	71	463.7	35.8	96	481.8	21.7
Passage Comprehension	Second	12	95	470.9	22.3	71	487	17.7	96	505.8	17.6
Calculation	First	9	93	451.5	33.5	71	466.4	40.9	96	474	23.3
Calculation	Second	12	94	478.8	20.8	71	494.9	20.5	96	508.1	20.4
Applied Problems	First	9	93	458.4	29.8	71	467.2	27.9	96	481	18.5
Applied Problems	Second	12	92	479.8	26.9	71	486.5	16.9	96	506.5	17.3
Spelling	First	9	86	447.9	27.7	71	460	31.7	96	479.6	23.3
Spelling	Second	12	73	468.2	25.0	71	481.1	23.5	96	508.8	21.5
Writing Samples	First	9	93	461.5	48.9	71	468.4	37.9	96	481.4	24.0
Writing Samples	Second	12	91	490.5	27.3	71	489.4	16.8	96	502.9	18.4

Table 3

Standardized Effect Sizes for Between-Sample Differences in W Scores from the Woodcock-Johnson IV (WJ IV) Tests of Achievement

Test	Measurement/ Age Grouping	*WJ IV Clinical Sample Compared to WJ IV Normative Sample*	*Study Sample Compared to WJ IV Normative Sample*	*Study Sample Compared to WJ IV Clinical Sample*
Word Attack	First	-.787†	-.839‡	-.017†
Word Attack	Second	-1.575†	-1.077†	.304†
Letter-Word Identification	First	-.793‡	-1.520‡	-.470†
Letter-Word Identification	Second	-1.198‡	-1.485‡	-.420‡
Passage Comprehension	First	-.834‡	-1.717‡	-.882‡
Passage Comprehension	Second	-1.066†	-1.737†	-.787†
Calculation	First	-.326‡	-.963‡	-.401†
Calculation	Second	-.645†	-1.437‡	-.778†
Applied Problems	First	-.748‡	-1.224‡	-.304†
Applied Problems	Second	-1.171†	-1.188†	-.289†
Spelling	First	-.839‡	-1.358‡	-.409†
Spelling	Second	-1.236†	-1.758†	-.532†
Writing Samples	First	-.422†	-.519†	-.156†
Writing Samples	Second	-.759†	-.669‡	.050†

Note. †Sample sizes are unequal, Hedge's *g* is reported. ‡There is a large difference in the standard deviations for these samples, Glass's delta is reported using the standard deviation for the WJ IV normative sample as a standardizer.

Table 4

Mean W Difference Scores and Relative Proficiency Index Scores for the Study Sample

Cluster	Before/During Pandemic	Mean *W* Difference Score	Mean RPI Score
Word Attack	Before	-15.1	60.7
	During	-18.2	54.1
Letter-Word Identification	Before	-44.4	22.1
	During	-39.7	29.2
Passage Comprehension	Before	-38.8	23.4
	During	-34.8	28.1
Calculation	Before	-23.5	50.8
	During	-30.3	33.5
Applied Problems	Before	-23.5	44.6
	During	-26.7	41.8
Spelling	Before	-32.0	30.2
	During	-41.7	21.5
Writing Samples	Before	-22.8	60.0
	During	-12.4	70.1

Notes. *W* Difference scores represent the distance of an examinee's *W* score from a grade-appropriate reference point (i.e., average scores at the examinee's grade level based on WJ IV normative data). A negative score indicates performance below the reference point. RPI = Relative Proficiency Index. The average range for this index is 82 to 95. Scores ranging from 67 to 82 are indicative of limited to average proficiency. Scores ranging from 24 to 67 indicate limited proficiency and scores below 24 indicate that proficiency is very limited.